Towards an On-Line Analysis of Tweets Processing

Sandra Bringay, Nicolas Béchet, Flavien Bouillot, Pascal Poncelet, Mathieu Roche, Maguelonne Teisseire

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

Towards an On-Line Analysis of Tweets Processing

Sandra Bringay[1,2], Nicolas Béchet[3], Flavien Bouillot[1],
Pascal Poncelet[1], Mathieu Roche[1], and Maguelonne Teisseire[1,4]

[1] LIRMM – CNRS, Univ. Montpellier 2, France
{bringay,bouillot,poncelet,mroche}@lirmm.fr
[2] Dept MIAp, Univ. Montpellier 3, France
[3] INRIA Rocquencourt - Domaine de Voluceau, France – nicolas.bechet@inria.fr
[4] CEMAGREF – UMR TETIS, France – maguelonne.teisseire@cemagref.fr

Abstract. Tweets exchanged over the Internet represent an important source of information, even if their characteristics make them difficult to analyze (a maximum of 140 characters, etc.). In this paper, we define a data warehouse model to analyze large volumes of tweets by proposing measures relevant in the context of knowledge discovery. The use of data warehouses as a tool for the storage and analysis of textual documents is not new but current measures are not well-suited to the specificities of the manipulated data. We also propose a new way for extracting the context of a concept in a hierarchy. Experiments carried out on real data underline the relevance of our proposal.

1 Introduction

In recent years, the development of social and collaborative Web 2.0 has given users a more active role in collaborative networks. Blogs to share one's diary, RSS news to track the lastest information on a specific topic, and tweets to publish one's actions, are now extremely widespread. Easy to create and manage, these tools are used by Internet users, businesses or other organizations to distribute information about themselves. This data creates unexpected applications in terms of decision-making. Indeed, decision makers can use these large volumes of information as new resources to automatically extract useful information.

Since its introduction in 2006, the Twitter website [5] has developed to such an extent that it is currently ranked as the 10^{th} most visited site in the world [6]. Twitter is a platform of microblogging. This means that it is a system for sharing information where users can either follow other users who post short messages or be followed themselves. In January 2010, the number of exchanged tweets reached 1.2 billion and more than 40 million tweets are exchanged per day [7]. When a user follows a person, the user receives all messages from this person, and

[5] http://twitter.com
[6] http://www.alexa.com/siteinfo/twitter.com
[7] http://blog.twitter.com/2010/02/measuring-tweets.html

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

in turn, when that user tweets, all his followers will receive the messages. Tweets are associated with meta-information that cannot be included in messages (e.g., date, location, etc.) or included in the message in the form of tags having a special meaning. Tweets can be represented in a multidimensional way by taking into account all this meta-information as well as associated temporal relations. In this paper, we focus on the datawarehouse [1] as a tool for the storage and analysis of multidimensional and historized data. It thus becomes possible to manipulate a set of indicators (measures) according to different dimensions which may be provided with one or more hierarchies. Associated operators (e.g., Roll-up, Drill-down, etc.) allow an intuitive navigation on different levels of the hierarchy.

This paper deals with different operators to identify trends, the top-k most significant words over a period of time, the most representative of a city or country, for a certain month, in a year, etc. as well as the impact of hierarchies on these operators. We propose an adapted measure, called $TF\text{-}IDF_{adaptive}$, which identifies the most significant words according to level hierarchies of the cube (e.g., on the location dimension). The use of hierarchies to manage words in the tweets enables us to offer a contextualization in order to better understand the content. We illustrate our proposal by using the MeSH [8] (Medical Subject Headings) which is used for indexing PubMed articles [9].

The rest of the paper is organized as follows. Section 2 describes a data model for cubes of tweets and details the proposed measure. In Section 3, we consider that a hierarchy on the words in tweets is available and propose a new approach to contextualize the words in this hierarchy. We present some results of conducted experiments in Section 4. Before concluding by presenting future work, we propose a state-of-the-art in Section 5.

2 What is the most appropriate measure for tweets?

2.1 Preliminary Definitions

In this section we introduce the model adapted to a cube of tweets. According to [2], a fact table F is defined on the schema $D = \{T, \ldots, T_n, M\}$ where T_i $(i = 1, .., n)$ are the dimensions and M stands for a measure. Each dimension T_i is defined over a domain D partitioned into a set of categories C_j. We thus have $D = \cup_j C_j$. D is also provided with a partial order $\sqsubseteq_D$ to compare the values of the domain D. Each category represents the values associated with a level of granularity. We note $e \in D$ to specify that e is a value of the dimension D if there is a category $C_j \subseteq D$ such that $e \in \cup_j C_j$. Note that two special categories are distinguished and are present on all dimensions: $\perp_D$ et $\top_D \in C_D$ corresponding respectively to the level of finer and higher granularity. In our approach, the partial order defined on the domains of the dimensions stands for the inclusion of keywords associated to the values of the dimensions. Thus, let $e_1, e_2 \in \cup_j C_j$ be two values, we have $e_1 \sqsubseteq_D e_2$ if e_1 is logically contained in e_2.

[8] http://www.ncbi.nlm.nih.gov/mesh
[9] http://www.ncbi.nlm.nih.gov/PubMed/

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

2.2 The data model

We instantiate the data model of the previous section to take into account the different dimensions of description and propose a specific dimension to the words associated to tweets.

Let us consider, for example, the analysis of tweets dedicated to the Duncan diet (e.g., "*The Dukan diet is the best diet ever, FACT!!! Its just meat for me for the next 5 day YEESSS*"). We wish, for example, to follow the comments or opinons of different people on the effectiveness of a diet. In order to extract the tweets, we query Twitter using a set of seed words: *Duncan, diet, slim, protein, etc.*. In this case, the original values of the word dimension are $dom(word) = \{$ *Duncan, diet, slim, protein, ...*$\}$.

Figure 1 illustrates the data model. We find the dimension ($location \perp_{location} = City \leq State \leq Country \leq \top_{location}$), and the dimension $time$ ($\perp_{time} = day \leq month \leq semester \leq year \leq \top_{time}$).

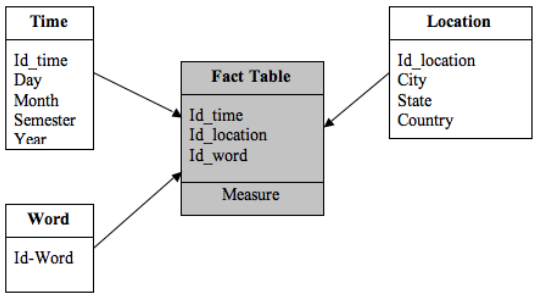

Fig. 1. The schema related to a diet application

The domain of the *word* dimension is that of the seed words with the words appearing frequently with them. In the fact table, several measures may be used. Traditionally it is the TF-IDF. This issue is addressed in the next section.

2.3 Towards an appropriate measure

Relying only on knowledge of the hierarchy in a cube does not always allow a good aggregation (i.e., corresponding to a real situation). For instance, the characteristics of the words in tweets are not necessarily the same in a State and in a City. The aggregation measure that we propose is based on approaches from Information Retrieval (IR).

In our process, the first step is to merge the number of occurrences of words specific to a level. More precisely, we list all the words located in tweets that match a given level (e.g., City, State, Country). If the user wishes to focus the search on a specific City, the words from the tweets coming from this city form a vector. We can apply this same principle to the State by using a Roll-up operator. The aim of our work is to highlight the discriminant words for each level.

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

Traditionally, $TF\text{-}IDF$ measure gives greater weight to the discriminant words of a document [3]. Like [4], we propose a measure called $TF\text{-}IDF_{adaptative}$ aiming to rank the words according to the level where the user is located and defined as follows:

$$TF_{i,j} - IDF_i^k = \frac{n_{i,j}}{\sum_k n_{k,j}} \times log_2 \frac{|E^k|}{|\{e_j^k : t_i \in e_j^k\}|} \quad (1)$$

where $|E^k|$ stands for the total number of elements of type k (in our example, $k = \{City, State, Country\}$) which corresponds to the level of the hierarchy that the decision maker wants to aggregate. $|\{e_j : t_i \in e_j\}|$ is relative to the number of elements of type k where the term t_i appears.

3 A hierarchy of words for tweets

In this section, we adopt a hierarchy on the words to allow the contextualization of words in tweets.

3.1 The data and the model

For the hierarchy, we use the MeSH (Medical Subject Headings)[10] National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a twelve-level hierarchy that permits the search to be carried out at various levels of specificity. At the most general level of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders". More specific headings are found at more narrow levels, such as "Ankle" and "Conduct Disorder". In 2011, 26,142 descriptors are available in MeSH. There are also over 177,000 entry terms that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid".

The data model is updated to take into account this new dimension. Compared to the previous model (See Figure 1) the dimension "Word" has been replaced by MeSHWord. MeSHWord has a partial order, $\sqsubseteq_{MeSHWord}$, to compare the different values of the domain. One of the main problems with the use of this thesaurus is that different terms may occur at various levels in the hierarchy. This ambiguity raises the problem of using operators like Roll-up or Drill-down to navigate in the cube. In order to illustrate this problem let us consider the following example.

Example 1 *Let us consider the following tweet: "pneumonia & serious nerve problems. can't stand up. possible myasthenia gravis treatable with meds.". If we look in MeSH for the descriptor pneumonia, we find this term occurring in several places (See Figure 2). Depending on the position in the hierarchy, a Roll-up operation on pneunomia will not give the same result (i.e., "respiratory tract diseases" versus. "lung diseases").*

[10] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

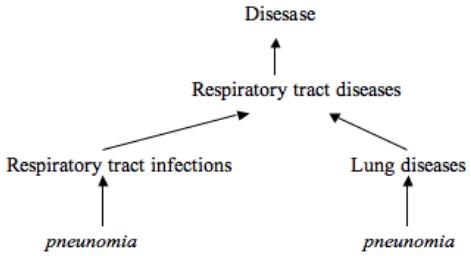

Fig. 2. An example of the MeSH thesaurus

3.2 How to identify the context of a tweet?

We have shown in Example 1, the difficulty of locating the context of a term in the hierarchy. However, a closer look at the tweet shows that the words themselves can be helpful to determine the context. In order to disambiguate polysemous words in the hierarchy of MeSH, we adapt the $AcroDef_{MI^3}$ method described in [5] where the authors show the efficiency of this method in a biomedical context. This measure is based on the Cubic Mutual Information [6] that enhances the impact of frequent co-occurrences of two words in a given context. For a context C, $AcroDef_{MI^3}$ is defined as follows:

$$AcroDef^{C}_{MI^3}(m1, m2) = \frac{(nb(m1\ and\ m2\ and\ C))^3}{nb(m1\ and\ C) \times nb(m2\ and\ C)} \quad (2)$$

In our case, we want to calculate the dependence between a word m to disambiguate and different words m_t of the tweets using the context of the hierarchy (i.e., parents p of the word m).

Example 2 *Let us consider the word 'pneunomia' to disambiguate in Example 1. Here we calculate the dependence between this word m and the other words following 'pneumonia' (nouns, verbs, and adjectives are selected with a Part-of-Speech process): 'serious' and 'nerve'. This dependence is calculated regarding the context of both possible fathers in the MeSH hierarchy. In order to predict where in the MeSH thesaurus we have to associate the word 'pneunomia', we perform the following operations:*

- $nb(pneunomia, m_t, "lung\ diseases") = 227$ *(number of returned pages with the queries 'pneunomia serious "lung diseases"' and 'pneunomia nerve "lung diseases"')*
- $nb(pneunomia, m_t, "respiratory\ tract\ infections") = 496$

The dependence of the terms is given by:

- $AcroDef^{"lung\ diseases"}_{MI^3}(pneunomia, m_t) = 0.02$
- $AcroDef^{"respiratory\ tract\ infections"}_{MI^3}(pneunomia, m_t) = 0.11$

Thus, in the tweet from Example 1, for the word pneunomia, we will preferably do the aggregation at the level of the concept 'respiratory tract infections' of the MeSH.

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

Note that this step of disambiguation, which is essential for data from MeSH, is quite costly in terms of the number of queries. It therefore seems more appropriate to call these functions during the ETL process rather than carrying out such processing when browsing the cube.

4 Experiments

In order to evaluate our approach, several experiments were conducted. These were performed using PostgreSQL 8.4 with the Pentaho Mondrian 3.20 environment. To extract the tweets related to the vocabulary used in MeSH, we focus on the tweets related to "Disease" and queries Twitter by using all the terms of the corresponding hierarchy. We collected 1,801,310 tweets in English from January 2011 to February 2011. In these experiments, we analyze the first words returned by the TF-IDF$_{adaptive}$ (highest scores). For example, the following table presents the first 12 words of tweets in the United States, for the State of Illinois and the City of Chicago during the month of January 2011.

United Sates	Illinois	Chicago
wart	risk	risk
pneumonia	vaccination	wart
vaccination	wart	pneumonia
risk	pneumonia	wood
lymphoma	wood	colonoscopy
common cold	colonoscopy	x-ray
disease	x-ray	death
meningitis	encephalitis	school
infection	death	vaccination
vaccine	school	eye infection
life	eye infection	patient
hepatitis	man	russia

Now we consider an example of the application of our approach. Figures *3* and *4* visualize the worldwide coverage of the words *hepatitis* and *pneunomia* excluding the United States, the United Kingdom, and Canada. This coverage is obtained by fixing the location dimension and by examining the frequency of the Word over the period.

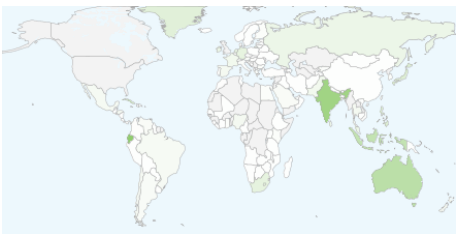

Fig. 3. Distribution of the use of the word hepatitis

Fig. 4. Distribution of the use of the word pneunomia

Finally we evaluated the prediction measure (i.e., $AcroDef_{MI^3}$) within the MeSH hierarchy (see section 3.2). We extracted more than 5,000 Facebook messages (the same kind of messages as tweets) from the *food* topic. These messages

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

contain at least one polysemous term (i.e. a term which can be associated to the hierarchy *food and beverages*) and one or two other hierachies of MeSH: *Eukaryota*, *lipids*, *plant structures*, and so forth. A correct prediction means that $AcroDef_{MI^3}$ associates this polysemous term with the *food and beverages* concept. In the following table, three types of elements are used in order to characterize the hierarchy (context of the $AcroDef_{MI^3}$ measure): Father (F), grand-father (GF), and father + grand-father (FGF). This table shows that (1) the use of more generic information (grand-father) is more relevant, (2) the association of all the available information improves the quality of the prediction. In our future work we plan to add other hierarchical information (e.g. son, cousins).

Elements of the hierarchy used	F	GF	FGF
Prediction	60.8%	63.6%	68.0%

5 Related work

The analysis of textual data from tweets has recently been addressed in many research studies and many proposals exist. For example, in [7], the authors propose analyzing the content of the tweets in real time to detect alarms during an earthquake. The authors of TwitterMonitor [8] present a system to automatically extract trends in the stream of tweets. A quite similar approach is proposed in [9]. However, to the best of our knowledge, most existing studies mainly focus on a specific analysis of tweets and do not provide general tools for the decision maker (i.e., for manipulating the information embedded in tweets according to their needs). Thus, few studies have been interested in the use of cubes to manage tweets. Recent work has focused on integrating textual data in data warehouses. In this context, aggregation methods suitable for textual data have been proposed. For example, in [10], the authors propose using Natural Language Processing techniques to aggregate the words with the same root or the same lemmas. They also use existing semantic tools such as Wordnet or Roget to group words together. Apart from using morpho-syntactic and semantic knowledge, other studies consider numerical approaches from the field of Information Retrieval (IR) to aggregate textual data. Thus, the authors of [11] propose aggregating documents according to keywords by using a semantic hierarchy of words found in the datawarehouses and some measures from IR. Such methods from IR are also used in the work of [2] which takes into account a "context" and "relevance" dimension to build a datawarehouse of textual data called R-cube. Other approaches add a new specific dimension. For example, in [12], the authors add a "topic" dimension and apply the PLSA approach [13] to extract the themes representing the documents in this new dimension. Finally, in [14] the authors propose aggregating parts of documents to provide the decision maker with words specific to the aggregation. In this context, the authors use a first function to select the most significant words using the classical $TF\text{-}IDF$ measure.

Colloque : 22nd International Conference on Database and Expert Systems Applications, DEXA 2011, Toulouse, 29 August - 2 September 2011
Author-produced version of the article published in Lecture Notes in Computer Science, 2011, 6861(2), 154-161.
The original publication is available at http://www.springer.com
DOI: 10.1007/978-3-642-23091-2_15

6 Conclusion

In this paper we proposed a new approach to analyze tweets from their multidimensional characteristics. The originality of our proposal is to define and manipulate cubes of tweets. We have shown through two different models and applications: no predefined hierarchy on tweets (i.e., diet analysis) and existing hierarchy (i.e., using the MeSH thesaurus), that the analysis of tweets requires the definition of new measures and that a contextualization step is relevant.

Future work involves several issues. First we want to extend the proposed approach to take into account opinions or feelings expressed in the tweets. Recent studies analyze the mood of people (e.g., http://twittermood.org/). We want to enhance these approaches by analyzing the content of tweets and thus be able to automatically extract knowledge such as: who are the people who followed a diet and are dissatisfied? Secondly, we wish to consider tweets as available in the form of a stream and propose new techniques for efficiently storing the data.

References

1. Codd, E., Codd, S., Salley, C.: Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. In: White Paper. (1993)
2. Pérez-Martínez, J.M., Llavori, R.B., Cabo, M.J.A., Pedersen, T.B.: Contextualizing data warehouses with documents. Decision Support Systems **45**(1) (2008) 77–94
3. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11) (1975) 613–620
4. Grabs, T., Schek, H.J.: ETH Zurich at INEX: Flexible Information Retrieval from XML with PowerDB-XML. In: XML with PowerDB-XML. INEX Workshop, ERCIM Publications (2002) 141–148
5. Roche, M., Prince, V.: Managing the acronym/expansion identification process for text-mining applications. Int. J. of Software and Informatics **2**(2) (2008) 163–179
6. Daille, B.: Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. PhD thesis, Université Paris 7 (1994)
7. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: Proceedings of WWW. (2010) 851–860
8. Mathioudakis, M., Koudas, N.: Twittermonitor: trend detection over the twitter stream. In: Proceedings of SIGMOD, Demonstration. (2010) 1155–1158
9. Benhardus, J.: Streaming trend detection in twitter. In: National Science Foundation REU for Artificial Intelligence, NLP and IR. (2010)
10. Keith, S., Kaser, O., Lemire, D.: Analyzing large collections of electronic text using olap. Technical Report TR-05-001, UNBSJ CSAS (2005)
11. Lin, C.X., Ding, B., Han, J., Zhu, F., Zhao, B.: Text Cube: Computing IR Measures for Multidimensional Text Database Analysis. In: Proc. of ICDM. (2008) 905–910
12. Zhang, D., Zhai, C., Han, J.: Topic cube: Topic modeling for olap on multidimensional text databases. In: In Proc. of SIAM. (2009) 1123–1134
13. Hofmann, T.: Probabilistic latent semantic analysis. In: In Proc. of Uncertainty in Artificial Intelligence, UAI'99. (1999) 289–296
14. Pujolle, G., Ravat, F., Teste, O., Tournier, R.: Fonctions d'agrégation pour l'analyse en ligne (OLAP) de données textuelles. Fonctions TOP_KW et AVG_KW opérant sur des termes. Ingénierie des Systèmes d'Information **13**(6) (2008) 61–84